Research Article

International Journal of Advanced Computer Research, Vol 7(33)
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2017.733033

VBTones a visual to auditory device for the blind

Nooshin Riahi[1*], Seyede Fatemeh Mirhoseini [2] and Eyesun Mehrbani[2]
Department of Computer Engineering, Alzahra University, Tehran, Iran[1]
Department of Electronic Engineering, Alzahra University, Tehran, Iran[2]

Received: 21-September-2017; Revised: 08-November-2017; Accepted: 09-November-2017

Abstract
Sensory substitution contains methods which send perceptible information of a sensory organ to the brain through other sensory modalities, resulting in the rehabilitation of the lost perception. Additionally, they may also be taken into consideration as research devices to examine the brain mechanisms and its cross-modal function. Visual to auditory sensory substitution used in experimentation aimed to provide visually impaired people with environmental perception. According to this paper, vOICe based tones (VBTones) has been introduced as a tool converting visual data into auditory output, where each image row is assigned to a specific sound frequency and the sound amplitude is referred to brightness of the pixel, where the image is scanned, converted the sound is generated column by column. Gradually forming a continuous sound made of single column waves which are the sum of multiple sinusoidal waves of different amplitudes and frequencies related to specific pixels. The objective of this study is to maximize the perception of visual information through audio. Thus, elimination of unnecessary image information is required. By applying anisotropic filtering methods in addition to a Laplacian-Gaussian filter, the produced sound turned out to be finer in texture and more perceivable. Moreover, this study shows analysis of the efficiency of this tool and several enhancements on both sighted participants and the visually impaired ones. The blind scored 87% and the sighted 78% accuracy in recognition in the designed test.

Keywords
Sensory substitution, Image to audio conversion, Blind rehabilitation, Anisotropic filtering.

1.Introduction
Destruction in an organ frequently happens between human beings and sets many obstacles in their lives. Although after a while recovery and adoption appear, many aids and methods are developed to compensate the lost perception. A trade called sensory substitution declares that information can be transmitted to the brain through different paths and methods leading to environmental recognition [1] with regard to flexibility and plasticity and neuroplasticity of the brain [2]. Plasticity of the brain refers to the capability of combining information from different sensory modalities [3]. Sensory substitution devices (SSDs) provide a possibility to convert visionary data to tactile or auditory one [4]. Braille was one of the primary facilities for this aim. Furthermore, developed visual to tactile devices were experimenting at the late 1960s [5]. However, on account of portability issues, high energy consumption and skin reactions, visual-tactile SSDs are more problematic [6, 7].

Yet, they are under research to be developed using more efficient actuators [8].

Scientific fundamental and backgrounds of these devices are rely on the unidirectional assignment of perceptible features from one sense to the other one. In this operation, data processing is equivalent to encoding the parameters of the first disabled, sensory organ of information that can be perceived by the other one. Despite many breakthroughs in this field, complex conversions and combinations pose a challenge to the performance of these structures. The main reasons for the SSDs not to be widespread are insufficient facility for using in the real world, lack of organized training experiments and interfering with sight restoration efforts by altering the original functions of visual cortex [2].

This study aims to provide a visual to auditory tool for the blind considering vOICe algorithms. Moreover, its effect is studied on both sighted participants and visually impaired users. Consequently, some corrections are applied to enhance performance. There is an expectation that

*Author for correspondence

the blind group would be able to outperform the sighted in recognition during the test [9, 10].

Primary attempts in using visual to audio converters were done in 1998 by Capelle et al. considering the hearing parameters such as the just noticeable difference (JND) and pleasant distances. Eliminating the resonance effect, sounds are chosen from 50-15000 Hz and the sound assigned to each pixel is multiplication of basic frequency (approximately 50-60 Hz) depending on pixel number and an exponential function [2].

The prosthesis is substituting vision by audition (PSVA) [2], vOICe [4] and EyeMusic [6, 7, 1] are the well-known devices in vision to the auditory conversion field which contain a set of camera and speaker as a connection to the environment. *Figure 1* is an illustration of this connection. In PSVA, each pixel refers to a specific sinusoidal wave frequency and the pixel luminance is assigned to the sound loudness [2]. While, vOICe and EyeMusic encode vertical coordination to individual frequencies [4, 6, 7]. The image in vOICe is gray scaled, whereas EyeMusic can demonstrate colors by using musical instruments such as Reggae Organ, Rapman's Reed, choir, string and Brass instruments in pentatonic scales. For its distinguishable intervals and prevention from resonance effect [6, 7]. vOICe ceiling frequency is 5000Hz. In contrast, the frequency in EyeMusic is limited to 1,568Hz in light of the fact that sounds with high energy in the range of 2500-5000 Hz can sound unpleasant [11].

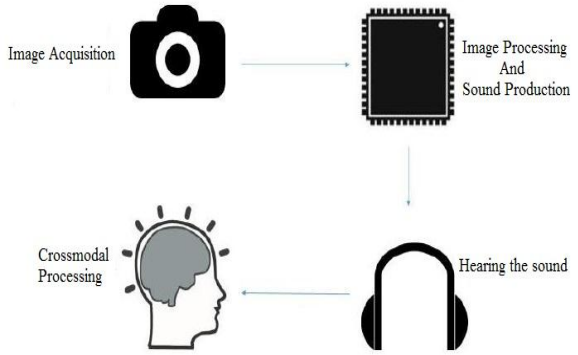

Figure 1 General overview of sensory substitution prototype

In these devices, the image is scanned column by column. Therefore, soundscape, of each column is presented sequentially. In a nutshell, the main differences between these tools are classified as an experimental training time, speed of data processing in the brain and the similarity between object recognition in the real environment and the simulation. In the environmental use of the device is involved with depth, motion and color perception. Considering high resolution, ease of training procedure, fast processing for dynamic objects and ability of depth perception, vOICe takes precedence over the other [4].

In vOICe algorithm, not only the color parameter has been eliminated to avoid overwhelming the brain by music assignment processing, but it can also provide depth and motion perception. Hence, it takes priority over similar tools [4, 6, 7].

Despite of a vast majority of neurological and psychological relevant information, lack of sufficient researches and resources related to the structure of these devices is observed. This paper discusses the application of a new filtering method based on vOICe algorithm with a different set of sinusoidal sound wave frequencies. Several cases took the designed test and the results were demonstrated.

In section one of these paper relevant researches is discussed and the sensory substitution approach and visual to auditory devices for the visually impaired users are introduced. Section two represents the algorithm used for image to audio conversion in the subsections. The experiments described and results are presented in sections three and four respectively. Section five discusses this study benefits and a comparison to the previous literature. The sixth section puts forward conclusions and future attempts.

2.Methodology

Sensory substitution, vision to auditory conversion for the visually impaired users Conventionally, visually impaired users replace other senses, such as tactile and hearing instead of the visual sense. Incidentally, SSDs has been used in order to ease environmental recognition in accordance with the plasticity of the brain and multifunctional specialization.

Various transforms and assignments in auditory sensation led to visual substitution. For instance, PSVA and vOICe encode the brightness and pixel coordinate to the volume and frequency of sound respectively. PSVA algorithm due to harsh in locating the objects in the field failed to develop. Because PSVA assigned higher frequencies to the pixels in a row as they approached the right side of the image. Thus, pixels around the right side of a row

had a similar sound to lower rows in the left side. EyeMusic and vOICe are still under research, considering the poor performance of both devices in 3D use. Additionally, there is a new device which can assign colors and shape to auditory and tactile sense respectively, where it is not applied in the real world yet [8].

Because of the disorder in the function of a sensory organ, data acquisition and processing are carried out by other sensory organ. In vision to auditory substitution, in order to make an image perceivable for the ears, camera can be utilized instead of eyes taking the image and transferring it to the processor.

Figure 2 shows the image to audio conversion. We used the pure sinusoidal waveforms similar to what are used in vOICe algorithm. For most people hearing range falls rapidly at 4 KHz, reducing the frequency variation range of the device, enhances the sound in texture. Particularly, in this study the domain where the amplitude perception according to frequency variations remains constant is better used. Sinusoidal frequencies are assigned to individual rows, i.e. pixels of a row share the same frequency, where higher rows in the coordinate are assigned higher frequencies. The image is scanned column by column. Thus, multiple sine waves of the related pixels are accumulated and divided by the number of pixels, generating the sound. Creating this short duration sound (approximately 15mS for 1Hz capturing rate), brighter pixels are of more importance. Sounds are generated and played successively, ultimately forming the image sound.

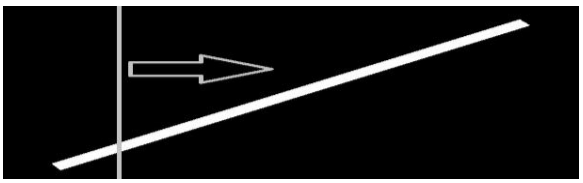

Figure 2 Left to right column by column swiping and real-time sound generation

Simulation process

Figure 3 shows the process steps of the tool. The steps described in the following.

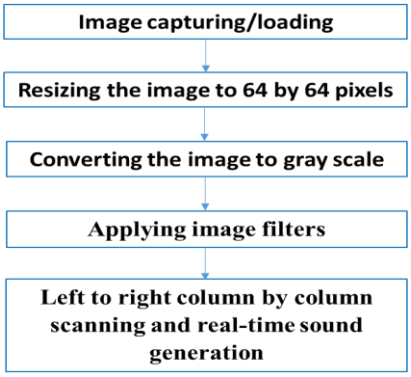

Figure 3 The process steps of the tool

2.1Image acquisition (capturing/loading)

The algorithm receives an image as input. The image can be either captured by a low quality camera or loaded from some storage. As far as the camera selection is considered, lower resolution is preferred. Because resizing the image by a lossy compression or even pulse with modulation over the digital camera frequency, adds up to the processing complexities in addition to memory consumption.

2.2Resizing the image to 64 by 64 pixels

Due to the limited number of distinguishable frequencies in the hearing range and desires to avoid resonance, the image must be compressed to have only 64 pixels in a column. This compression is obtained by calculating the mean value of adjacent pixels, thus it is considered lossy and irreversible. Although, higher compression in the image results in larger differences between individual frequencies (thus an easier distinction), data loss of the image domain will be inevitable. *Figure 4* shows the image degradation as the size decreases.

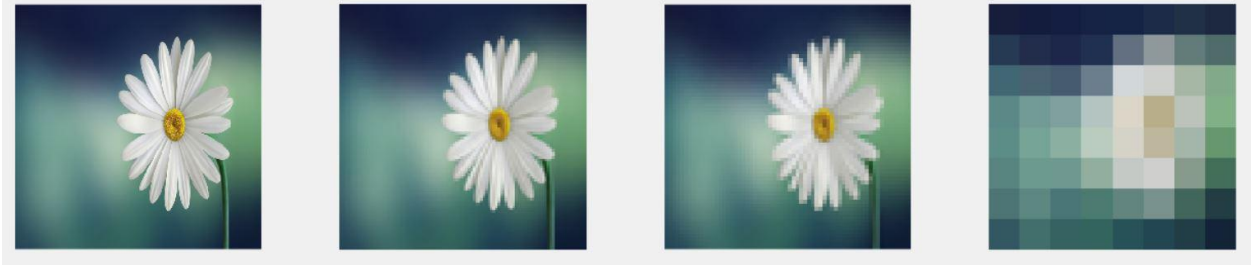

Figure 4 The image degradation as the size decreases for sizes from left to right; 512 by 512, 100 by 100, 64 by 64, 8 by 8

2.3Converting the image to grayscale (if required)

The RGB or YCbCr image is converted to grayscale image with 256 gray levels. Cameras with YCbCr output are preferred for easier extraction of grayscale data through Y channel. In converting the RGB image, calculation of the mean value between red, green and blue values is required to obtain the brightness level. The sinusoidal wave amplitudes are quantized by 256 levels in the audio domain. Thus, a linear mapping is established between image and audio domain which can be held in a byte of digital storage.

2.4 Applying image filters

To achieve the highest possible data perception rate per image, elimination of the redundant information is required. This elimination enhances the audio texture in understanding and the mapping process. Area with very low image frequency (low variation rate) or very high frequencies (e.g. the edges) are to be omitted. For instance, in the image of a tree, small leaves and branches are assumed to be redundant because when accumulated together, none of them will be individually noticed. Low frequency parts of the trunk are also redundant for some important image features could be lost by the human auditory perception, should the trunk not being eliminated from the image; since it produces loud noise like sound. The word elimination here refers to reducing the brightness and thereby the corresponding loudness of the produced sound. Finally, higher image frequencies are darkened and so for the very low frequencies. It can be approximately considered a band pass filter. *Figure 5* shows a tree in the filtering process. *Figure 6* shows left and right isotropic filtering regarding the edges.

In the implemented version, the mentioned amplitude reduction is under focus by means of anisotropic diffusion. In anisotropic filtering methods, unlike the Gaussian filters, fading effect is not applied homogeneously over all the image parts. The filter equation is very similar to the heat equation, where the high image frequency parts are corresponding to thermal insulators. Sharper edges are more resistant to the fading effect. Additionally, this type of diffusion has proved to be useful for noise removal applications [12]. Ultimately, this method will result in an enhanced filtering for elimination of the redundant information [13].

Figure 5 Left: The original image. Right: The filtered image

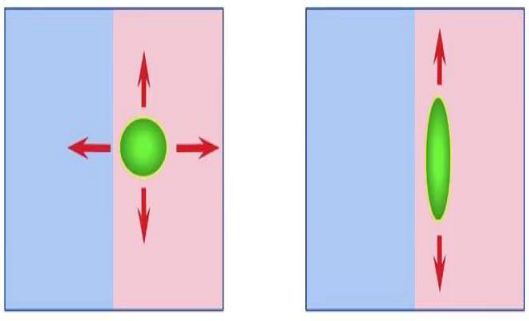

Figure 6 Left: isotropic filtering, Right: anisotropic filtering regarding to the edges

The equations for isotropic filtering are commonly formed like equation one:

$$\eth I(x,y,t)/\eth t = \Delta I \quad (1)$$

Anisotropic filtering is a method of enhancing the image quality by reducing detail in one direction. While, anisotropic filter treats all axes equally.

Anisotropic filtering equation is shown below.

$$\eth I(x,y,t)/\eth t = \text{div}\,(g\,|\nabla I|\,\nabla I) \quad (2)$$

Where Δ denotes Laplacian, ∇ gradient, div divergent and I(x,y,t) is the diffusion coefficient that controls the rate of diffusion [14]. The second order divergent

basically stops the diffusing wave from passing through sharp edges.

In the next step, the low frequency redundancies are omitted using a Laplacian-Gaussian (edge detector) filter. This filtering sequence of anisotropic and Laplacian-Gaussian is set to avoid extra added edges that might be caused by the edge detector over though surfaces of the image. For the edge detector that deletes low variations, may bold and even add some unimportant and invisible edges through which the anisotropic filter cannot pass. The final image is quite similar to a first sketch, shaped only by the outer lines. As a result, fewer sinusoids are mixed and individual frequencies can be understood better, helping the user to comprehend more complex configurations and structures. Step by step filter application is presented in *Figure 7*. Discrete cosine transform (DCT) is applied to the last three parts of the image presented in *Figure 8*. As shown in the first part of *Figure 8*, the energy is spread over a large domain of frequencies. By applying the anisotropic filter, variations in higher frequency sinusoids are reduced and the main energy is concentrated around the transform origin. In the third part, higher frequencies are bolded in the transform domain which is an expected result due to the edge detector application. This result certifies the purpose of anisotropic filtering; a main object out of background extraction.

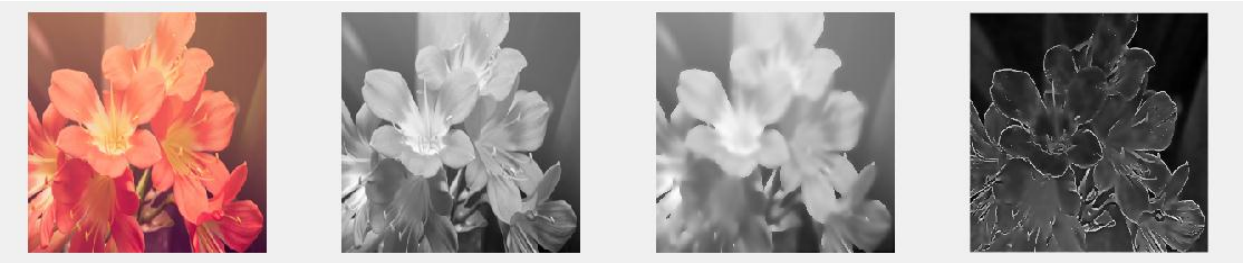

Figure 7 Step by step filtering of the image from left to right; original image, grayscale image, anisotropic applied image, Laplacian-Gaussian filter applied image

Figure 8 16X16 DCT of the images from left to right; grayscale image, anisotropic filtered image, anisotropic in addition to Laplacian-Gaussian filtered image

2.5Left to right column by column scanning and real-time sound generation

Transfer of the audio domain is performed based on 64 sinusoidal waves, each assigned to one of 64 pixels of a column. The waveforms are of different frequencies between 500 Hz to 5000 Hz. Because the amplitude perception according to frequency remains approximately constant over this domain. For most of the music and speech related sounds being in the mentioned frequency domain, the human ear is more familiar. As the 64 frequencies each relates to rows which are to be recognized, they must be selected according to Weber-Fechner law [15] in which JND must increase exponentially as the frequency rises linearly. To calculate the corresponding frequencies of each row, the following equation is used.

Frequency calculation according to height as shown below.

$$y(n) = 464.7\exp(0.074n) \quad (3)$$

Where y (n) is the frequency of the n'th pixel in a column calculated by an exponential curve fitting between 500 to 5000 Hz. For the m'th column the

total sound S(m) is generated by the following equation.

$$S(m) = 1/64 \sum_{n=1}^{64} (\mathrm{A(n)} \sin 2\pi y(n)) \qquad (4)$$

Where A(n) is a number between 0 and 1 corresponding to one of 256 gray scales, linearly mapped by a division in 256. Finally, the total image, sound is a series of column sounds, generated and played as the scanning process covers the capture from left to right. The total process is visualized in *Figure 9*.

3.Experiments

Primary testing included 43 black and white images with simple configurations. The goal was set on learning basic image perception concepts like thickness, height, location, and configuration perception. Several instances are shown in *Figure 10*. In the testing process, one sound is frequently played and the image is to be recognized. The tool environment is illustrated in *Figure 11* that includes the image and scanning speed, repetition and image options.

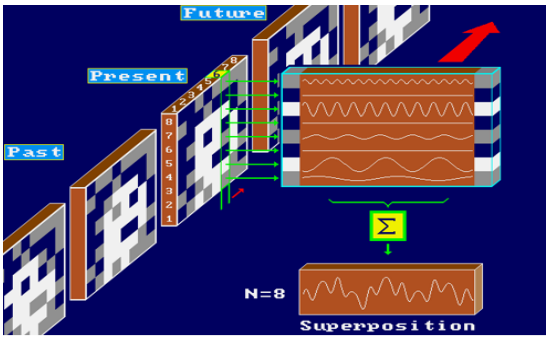

Figure 9 Image to audio conversion

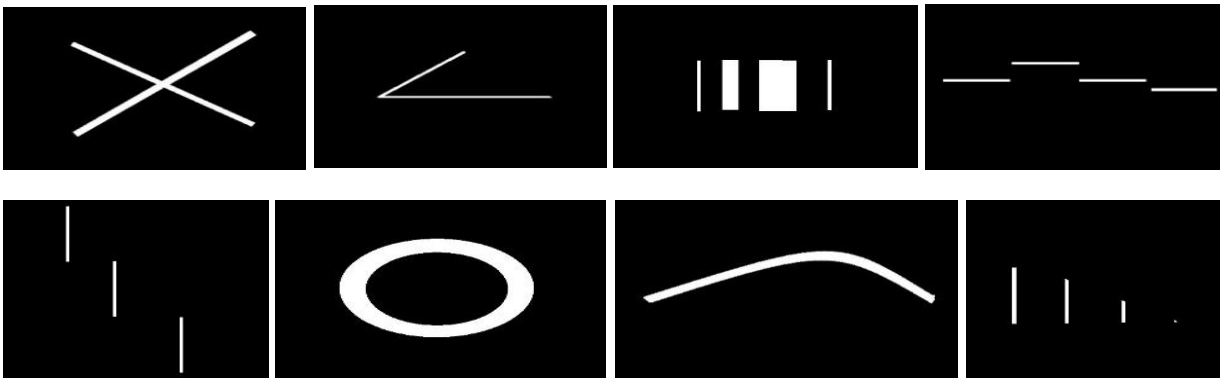

Figure 10 Examples of training images considering concepts of visual domain

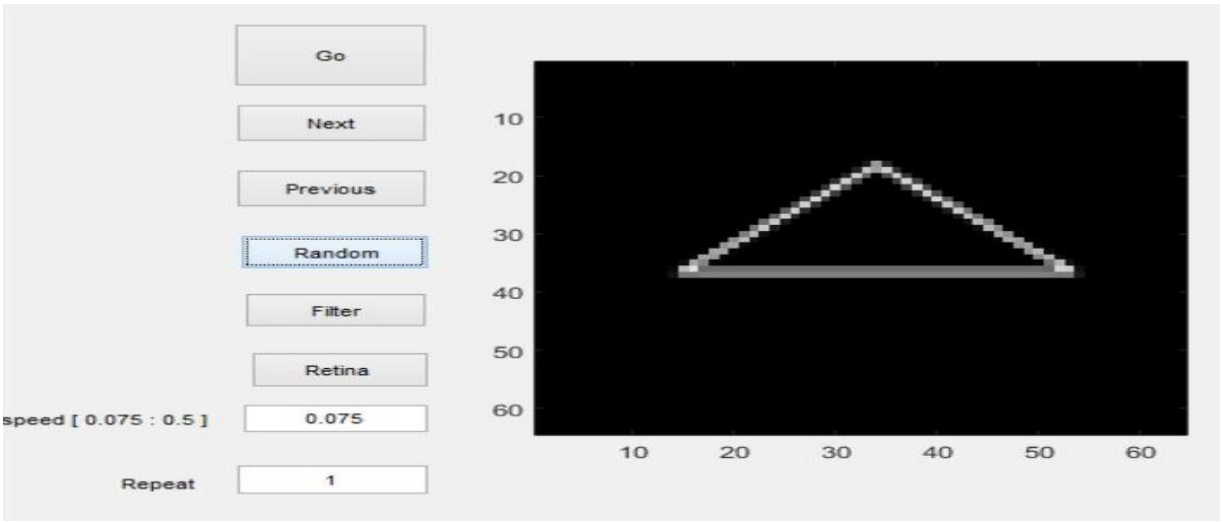

Figure 11 Tool environment

4.Experimental results

The candidate group included 13 blindfolded sighted people the test with on average 5 times repeat per capture. To accelerate the training time, in the primary examples was shown to the participant while the soundscapes were played [16]. For the blind group, the pictures were presented using embossed papers. The average value of training time was 52.1 minutes and 7.78 scores out of 10 was achieved. The results are presented in *Table 1*. The group also included 14 blind people under the test ten of which were able to finish the process. The test included 10 images with on average 4 times, repeat per image. The average value of training time was 44.5 minutes and 8.7 scores out of 10 was achieved. The results are presented in *Table 2*.

Table 1 The sighted

Gender	Age	Musical background	Training time	Score
Male	23	Yes	60	9
Male	22	Yes	60	10
Female	20	Yes	60	10
Female	17	Yes	120	10
Male	53	No	40	6
Female	8	No	45	8
Female	6	No	50	7
Male	11	No	50	7
Female	36	No	30	3
Female	22	No	40	8
Male	49	No	40	10
Female	17	No	30	8
Female	18	Yes	30	5

Table 2 The blind

Gender	Age	Musical background	Training time	Score
Male	17	Yes	40	10
Male	17	Yes	30	9
Male	18	Yes	40	9
Male	17	No	30	10
Male	17	No	30	10
Male	12	No	30	10
Male	18	No	40	6
Male	17	Yes	45	9
Male	18	Yes	60	7
Male	19	No	60	7
Male	17	Yes	40	10
Male	17	No	30	9
Male	18	Yes	40	9

The audio did not sound fine to 2 of the blind group. A middle-aged primary blind described the sounds bothersome and distracting. He also mentioned the lack of necessity of such devices and aids. The rest was interested and described the sounds appropriate.

It is observed that musical background has a remarkable effect on the ability of recognizing the frequency rise and falls and has enhanced the performance. The blind group with 7.6 minutes faster training achieved 0.92 better scores out of 10 due to the promotion in the remained senses when destruction occurs in one.

The late blind, got to a better insight of the algorithm. Their mental perception, as described by them, took place in two steps; hearing the sound and relating it to visual concepts, then describing the same card. However, the primary blind, had a more straight forward audio to image matching; without getting involved with the algorithm and definitions, relating the sounds to configurations. Thus, an intuition of light and darkness is expected by long term use of the device [4].

5.Discussion

Generally, this study justified the previous literature about the following subjects. First, audio-visual SSDs are a better choice rather than tactile ones in power consumption and portability issues [6, 7]. When compared to the experimented tactile based SSDs, auditory versions score a higher speed and accuracy which was achieved in this study [4, 5]. In the image processing domain anisotropic filtering improves the noise and redundancy removal. Finally, application of the serial filters has led to a smoother sound, i.e. a better distinctive audio. This will possibly be aiding the navigation capabilities and information gain in prolonged use. The experimental is in agreement with the blind group hearing sense outperformance in comparison with the sighted [9]. For the blindness enhances auditory obstacle circumvention [10]. Where, the training and test time was quite shorter in primer group. It is also acclaimed that the age of the participants takes negligible part in the final achievement. However, the congenital blind were fairly reluctant to use the provided SSD. Both groups achieved the abilities of recognizing the light intensity, shapes and figures, objects size and location.

6.Conclusion

This paper has presented a research on visual to auditory conversion algorithms which in addition to specific applications in rehabilitation of the blind, can be used as a tool for studying the cross modal performance of the brain in addition to the investigation the navigating capabilities based on the virtual environment [1, 6, 17, 18]. By applying anisotropic and Laplacian-Gaussian filters 87%

accuracy in the blind and 78% in the sighted was achieved. Due to the small size of statistic society the results are not quite trustworthy. Yet, the variance is fairly low.

A mentioned reason for such devices not to be widespread is the complexity of sights in the real world that leads to unrecognizable sounds. This necessitates more research on filtering the image, as well as audio domain adjustments. Better testing methods and more precise statistical analysis is required. Providing the proper hardware and renewing the previous design can improve the chance for better psychological, medical and statistical research.

Acknowledgment
The authors would like to thank Dr. H. Davoudi and Dr. M. Shokrizadeh for their technical advices and also management of Tehran educational department of disordered students for their good assistance during the experiments.

Conflicts of interest
The authors have no conflicts of interest to declare.

References
[1] Murphy MC, Nau AC, Fisher C, Kim SG, Schuman JS, Chan KC. Top-down influence on the visual cortex of the blind during sensory substitution. Neuroimage. 2016; 125:932-40.
[2] Capelle C, Trullemans C, Arno P, Veraart C. A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution. IEEE Transactions on Biomedical Engineering. 1998; 45(10):1279-93.
[3] Yang J, Li X, Li Q, Xiao X, Wu Q, Wu J. The effect of visual stimuli on auditory detection in an auditory attention task. In international conference on mechatronics and automation 2017 (pp. 1567-72). IEEE.
[4] Ward J, Meijer P. Visual experiences in the blind induced by an auditory sensory substitution device. Consciousness and Cognition. 2010; 19(1):492-500.
[5] Bach-y-Rita P, Kercel SW. Sensory substitution and the human–machine interface. Trends in Cognitive Sciences. 2003; 7(12):541-6.
[6] Abboud S, Hanassy S, Levy-Tzedek S, Maidenbaum S, Amedi A. EyeMusic: introducing a "visual" colorful experience for the blind using auditory sensory substitution. Restorative Neurology and Neuroscience. 2014; 32(2):247-57.
[7] Maidenbaum S, Abboud S, Amedi A. Sensory substitution: closing the gap between basic research and widespread practical visual rehabilitation. Neuroscience & Biobehavioral Reviews. 2014; 41:3-15.
[8] Hamilton-Fletcher G, Wright TD, Ward J. Cross-modal correspondences enhance performance on a colour-to-sound sensory substitution device. Multisensory Research. 2016; 29(4-5):337-63.
[9] King AJ. Crossmodal plasticity and hearing capabilities following blindness. Cell and Tissue Research. 2015; 361(1):295-300.
[10] Kolarik AJ, Scarfe AC, Moore BC, Pardhan S. Blindness enhances auditory obstacle circumvention: assessing echolocation, sensory substitution, and visual-based navigation. PloS One. 2017; 12(4):1-25.
[11] Kumar S, Forster HM, Bailey P, Griffiths TD. Mapping unpleasantness of sounds to their auditory representation. The Journal of the Acoustical Society of America. 2008; 124(6):3810-7.
[12] Rossovskii LE. Image filtering with the use of anisotropic diffusion. Computational Mathematics and Mathematical Physics. 2017; 57(3):401-8.
[13] Black MJ, Sapiro G, Marimont DH, Heeger D. Robust anisotropic diffusion. IEEE Transactions on Image Processing. 1998; 7(3):421-32.
[14] Positano V, Santarelli MF, Landini L, Benassi A. Nonlinear anisotropic filtering as a tool for SNR enhancement in cardiovascular MRI. In computers in cardiology 2000 (pp. 707-10). IEEE.
[15] Lanzara RG. Weber's law modeled by the mathematical description of a beam balance. Mathematical Biosciences. 1994; 122(1):89-94.
[16] Wright TD, Margolis A, Ward J. Using an auditory sensory substitution device to augment vision: evidence from eye movements. Experimental Brain Research. 2015; 233(3):851-60.
[17] Renier L, Collignon O, Poirier C, Tranduy D, Vanlierde A, Bol A, et al. Cross-modal activation of visual cortex during depth perception using auditory substitution of vision. Neuroimage. 2005; 26(2):573-80.
[18] Maidenbaum S, Buchs G, Abboud S, Lavi-Rotbain O, Amedi A. Perception of graphical virtual environments by blind users via sensory substitution. PloS One. 2016; 11(2):1-21.

Noushin Riahi received the B.S. degree from Isfahan University of Technology in 1986 and the M.S. and PhD. degree in Electrical and Electronics Engineering from Sharif University of Technology, Tehran, Iran in 1990 and 1998. Her research interests include speech and sound processing, opinion mining, text summarization, automatic machine translation, and biological signal processing. Currently, she is associate professor at Alzahra University, computer group. Topics taught by her include speech processing, advance computer architecture, digital signal processing, signals and systems theory, microprocessors, digital design, etc.
Email: n.riahi@alzahra.ac.ir

Seyede Fatemeh Mirhosseini is an undergraduate student of Electrical and Electronics Engineering at Alzahra University, Iran since 2013. She is currently working on signal processing applications in the blind rehabilitation. Her research interests are digital signal processing, neural network and machine learning.

Aysan Mehrbani is an undergraduate student of Electrical and Electronics Engineering at Alzahra University, Iran since 2014. She is currently working on signal processing applications in steganography and the blind rehabilitation. Her research interests are the machine vision and the application of neural networks in signal processing hardware and software development.